Development and individual differences in transitive reasoning

Bouwmeester, S.; Vermunt, J.K.; Sijtsma, K.

Published in:
Developmental Review

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com
ScienceDirect
Developmental Review 27 (2007) 41–74
DEVELOPMENTAL REVIEW
www.elsevier.com/locate/dr

Development and individual differences in transitive reasoning: A fuzzy trace theory approach

Samantha Bouwmeester [a,*], Jeroen K. Vermunt [b], Klaas Sijtsma [b]

[a] *Institute for Psychology, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands*
[b] *Tilburg University, The Netherlands*

Received 28 October 2005; revised 4 August 2006
Available online 17 October 2006

Abstract

Fuzzy trace theory explains why children do not have to use rules of logic or premise information to infer transitive relationships. Instead, memory of the premises and performance on transitivity tasks is explained by a verbatim ability and a gist ability. Until recently, the processes involved in transitive reasoning and memory of the premises were studied by comparing mean performance in fixed-age groups. In this study, an individual-difference model of fuzzy trace theory for transitive reasoning was formulated and tested on a sample ($N = 409$) of 4- to 13-year-old children. Tasks were used which differed with respect to presentation ordering and position ordering. From this individual-difference model expectations could be derived about the individual performance on memory and transitivity test-pairs.

The multilevel latent class model was used to fit the formalized individual-difference fuzzy trace theory to the sample data. The model was shown to fit the data to a large extent. The results showed that verbatim ability and gist ability drove the activation of verbatim and gist traces, respectively, and that children used combinations of these traces to solve memory tasks (testing memory of the premises) and transitivity tasks. Task format had a stronger effect on transitivity task performance than on memory of the premises. Development of gist ability was found to be faster than development of verbatim ability. Another important finding was that some children remembered the premise information correctly but were not able to infer the transitive relationship, even though the premises provided all the necessary information. This contradicts Trabasso's linear ordering theory which posits that memory of the premises is sufficient to infer transitive relationships.

* Corresponding author.
E-mail address: bouwmeester@fsw.eur.nl (S. Bouwmeester).

0273-2297/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.dr.2006.08.001

Keywords: Cognitive development; Cognitive modeling; Fuzzy trace theory; Individual differences; Transitive reasoning; Multilevel latent class analysis

Theory of transitive reasoning

General introduction

A transitive reasoning task requires the inference of an unknown relationship between two objects from the known relationships between each of these objects and a third object. For example, let three sticks, A, B, and C, differ in length, denoted as Y, such that $Y_A > Y_B > Y_C$; then given $Y_A > Y_B$ and $Y_B > Y_C$, the relationship between A and C can be inferred from these two relationships. In this example, the pairs $[A, B]$ and $[B, C]$ are the premise pairs and the relationships between the objects in the premise pairs constitute the premises.

A transitive reasoning task consists of a presentation stage and a test stage. At the presentation stage, the premise pairs are shown to the child. During the test stage, (s)he is asked to infer the transitive relationship from the premises; in the example, $Y_A > Y_C$. The object pair $[A, C]$ is the transitivity test-pair, because it tests the ability to infer a transitive relationship from the premises. The premise pairs may also be shown to test whether the child is able to remember the premises. When memory of the premises is tested, the premise pairs—in this example $[A, B]$ and $[B, C]$—are used as memory test-pairs.

According to Piaget (1947), children are capable of drawing transitive inferences when they understand the necessity of using rules of logic. When children know how to use these rules, they are able to solve any transitive relationship provided they can remember the premises. This understanding of rules of logic is acquired at the concrete operational stage, at approximately seven years of age. Then, for the first time children understand the reversibility principle (Piaget, 1942, 1947). A transitive inference effectively demonstrates this principle: When A is longer than B, the reversibility principle says that B must be shorter than A; and when one knows that A is longer than B, and C is shorter than B, one can use the reversibility principle to conclude that A is longer than C.

Children at the pre-operational stage—that is, at two to seven years of age (Piaget, 1947; see also Flavell, 1970)—do not yet understand the reversibility principle. Alternatively, these children consider objects or characteristics of objects in a nominal way, that is, not in relationship to other objects (Piaget, 1942). Due to this nominal thinking children are not capable of performing internalized operations on objects and they do not understand the necessity of using rules of logic. When a cue is provided about the ordering of the objects in a task, an understanding of such rules of logic may not be necessary to solve the task. For example, when all objects are presented simultaneously and when they are ordered on the dimension on which they differ, the position of the objects can be used for inferring their mutual relationships. Reasoning based on the use of cues is called functional reasoning. Functional reasoning is typical of the pre-operational stage. Piaget used transitive-reasoning tasks to study children's understanding of operational reasoning (Piaget, 1942; Piaget, Inhelder, & Szeminska, 1948).

According to Piaget's theory, memory of the premises is a necessary condition to solve a transitive relationship. Only when the premise information is available can children use the rules of logic to draw the transitive inference. Braine (1959) showed that after the premises had been learned children were able to draw transitive inferences at five years of age. He

argued that remembering the premises was the real problem for young children, not understanding rules of logic. Smedslund (1963, 1965, 1969) argued that Braine's results could be explained by a labelling strategy (this is a strategy in which a nominal label of an object is used to solve the task), but Trabasso and his colleagues (Riley, 1976; Riley & Trabasso, 1974; Trabasso & Riley, 1975; Trabasso, Riley, & Wilson, 1975) showed that four-year old children were able to draw a transitive inference on the test-pair [B, D] in a 5-object task ($Y_A < Y_B < Y_C < Y_D < Y_E$) in which they could not use the labelling strategy because B and D had no unique labels.

Bryant and Trabasso (1971) extensively trained 4- to 7-year-old children to remember the premises in a 5-object task, such as $Y_A < Y_B < Y_C < Y_D < Y_E$. On the basis of a series of experiments (Riley, 1976; Riley & Trabasso, 1974; Trabasso & Riley, 1975; Trabasso et al., 1975) Trabasso (1977) postulated his linear ordering theory (see also Bower, 1971), the basis of which is that children encode premises into an internal representation of the complete series. During training, reaction times and patterns of errors showed that children first learn the end-anchor premises of the series (i.e., $Y_A < Y_B$ and $Y_D < Y_E$) and then the premises in the middle (i.e., $Y_B < Y_C$, $Y_C < Y_D$). These findings agreed with both the distance hypothesis and the end-anchored hypothesis (Potts, 1972). Trabasso and his colleagues (Bryant & Trabasso, 1971; Trabasso, 1977; Trabasso et al., 1975) showed that children only need to remember the premises for being able to form an internal representation of the linear ordering. They concluded that an understanding of rules of logic is not necessary for transitive inference and that accurate memory of premises is sufficient for correct transitive inference. However, in contrast with these results, Halford and Galloway (1977), Grieve and Nesdale (1979) and Halford (1979) showed that children who remembered the premises sometimes failed to draw the transitive inference.

Both Piaget's and Trabasso's theories assumed that memory of the premises was a necessary condition for drawing correct inferences. This general agreement originated from the finding that children who responded correctly to transitivity test-pairs often justified their responses by restating the premises (Smedslund, 1963). Moreover, Trabasso and his associates (e.g. Trabasso, 1977; Trabasso et al., 1975), Adams (1978), and Perner, Steiner, and Staehelin (1981) showed that transitive inference was related to memory of the premises. However, Brainerd and Kingma (1984, 1985, see also Brainerd and Reyna, 1992) concluded that memory of the premises is *not* a prerequisite for transitive reasoning. Based on their own experiments and a re-analysis of studies by Halford and Galloway (1977), Kingma (1981), and Russell (1981) they showed that the solution of memory test-pairs and transitivity test-pairs are independent. Independence implies that the conditional probability of drawing a correct transitive inference (denoted T) given memory for the premises (denoted M), written as $P(T|M)$, is equal to the probability of drawing a correct transitive inference irrespective of memory of the premisses, written as $P(T)$; that is, $P(T|M) = P(T)$. Thus, two results are important here: children neither seemed to use rules of logic nor did they have to remember the premises to infer a transitive relationship. These findings contradicted both Piaget's and Trabasso's theory. Alternatively, Brainerd and Kingma explained these results by means of fuzzy trace theory.

According to fuzzy trace theory (Brainerd & Reyna, 1993, 2004; Reyna & Brainerd, 1990, 1992, 1995a, 1995b), incoming information is encoded and reduced to the essence. Representations of the same kind of information can be ordered by degree of exactness according to a hierarchy of gist (Reyna & Brainerd, 1990, 1995a). The different levels at which information is encoded are called *traces*. Traces that contain literal, well-articulated,

ornate representations preserving the content with exactitude are called *verbatim traces*. Traces that contain fuzzy, reduced, pattern-like information only holding the gist are called *gist traces*. Working memory holds both kinds of traces. Verbatim and gist traces are processed in parallel; that is, incoming information is processed simultaneously at different trace levels. Because the information stored in verbatim traces is fine-grained, retention is shorter than that of gist traces which do not require much working-memory capacity due to their impoverished structure (Reyna & Brainerd, 1995a).

Brainerd and Kingma (1984, 1985) and Brainerd and Reyna (1992; see also Reyna & Brainerd, 1990) posit that there is a pervasive inclination to thinking, reasoning and remembering by processing gist rather than verbatim traces. The mind picks up those traces that are suited for accurate performance on the presented task. According to Brainerd and Reyna (1990a; see also Brainerd & Reyna, 1990b) it is a natural habit of mind to process gist traces instead of verbatim traces, because the former has advantages over the latter with respect to trace availability, trace accessibility, trace malleability, and processing complexity. Tailored to transitive reasoning this means that instead of using the premises, more-global pattern information about the ordering of the objects is often sufficient to infer the transitive relationship (Brainerd & Kingma, 1984). Thus, when confronted with a 5-object task (e.g., $Y_A < Y_B < Y_C < Y_D < Y_E$) inference of the pattern "objects got smaller to the left" is enough to solve the transitivity test-pair $[B, D]$. This pattern information is processed in a gist trace. For memory test-pairs $[A, B]$, $[B, C]$, $[C, D]$, and $[D, E]$ children may use both verbatim and gist traces. That is, the premises are stored in verbatim traces. However, whenever a gist trace is available containing pattern information (e.g., "objects got smaller to the left") the memory test-pairs can also be inferred from a gist trace.

Brainerd and Kingma (1984) hypothesized three models to explain the parallel functioning of verbatim and gist traces. The unitary-trace model hypothesizes that, whenever a gist trace holding pattern information is available, children use this gist trace to solve both memory and transitivity test-pairs. The dual-trace model hypothesizes that children use verbatim traces when they solve memory test-pairs and gist traces when they solve transitivity test-pairs. The mixed-trace model hypothesizes that children solve memory test-pairs by means of verbatim traces when still in storage. When such information is not available anymore, they use gist traces.

Patterns of errors on memory test-pairs and transitivity test-pairs were analyzed to determine which of the three models best explains the parallel functioning of verbatim and gist traces. In agreement with Trabasso's linear ordering theory, Brainerd and Kingma (1984) showed that children performed better on the end-anchored test pairs of the spatial ordering than on the test pairs in between. This spatial-position effect was found both in the memory and transitivity test-pairs, suggesting that children used pattern (i.e., gist) information for both kinds of tasks. For the memory test-pairs, performance was the same irrespective of the order in which test pairs were presented. Thus, temporal-position effects, reflected by better performance on the premises presented first and last than on the premises in between, were absent. This result agreed with the unitary-trace model and disagreed with the mixed- and dual-trace models. Therefore, Brainerd and Kingma (1984) concluded that, when available, children use gist traces to solve both the memory and transitivity test-pairs.

For separate age groups, Brainerd and Kingma (1984) computed percentages of correct answers for memory and transitivity test-pairs. These percentages suggested that

performance on memory and transitivity test-pairs could be explained from the use of gist traces, and that the unitary-trace model fitted the data best. Moreover, older age groups on average showed better performance than younger age groups. However, by averaging results over children of the same age much variance in task performance remained unexplained: Notice that a spatial-position effect that is found at the group level does not imply a similar effect for each individual child. In fact, some children may indeed show a spatial-position effect while other children do not perceive any pattern at all which results in poor performance or, alternatively, they perceive the complete ordering which results in good performance on all memory and transitivity test-pairs. Theoretically, it is even possible that the average percentages-correct obscure that none of the individual children produced the pattern of task scores typical of a spatial-position effect. Thus, a theory that explains individual differences in task performance is badly needed.

Fuzzy trace theory offers a strong theoretical framework for explaining these individual differences, because it predicts individual differences in performance due to differences in the retrieval of verbatim and gist traces and the accuracy with which they are applied to solve tasks (Brainerd & Reyna, 1990a). Thus, the analysis of individual patterns of incorrect/correct scores on memory and transitivity test-pairs may shed more light on individual differences (see also Cooney, 1995).

The purpose of this study was to introduce an individual-difference model for fuzzy trace theory applied to transitive reasoning. This model was put to the test by means of multilevel latent class models that were fitted to data collected in a large representative sample of elementary-school students.

Individual-difference model of fuzzy trace theory applied to transitive reasoning

In our individual-difference model of fuzzy trace theory, children's performance on memory and transitivity test-pairs from a particular task is explained by the parallel retrieval and usage of verbatim and gist traces (Reyna & Brainerd, 1995a). Individual differences are explained from differences in the simultaneous use of verbatim and gist-trace levels. We assumed a verbatim ability and a gist ability on which children may differ. Tailored to transitive reasoning, verbatim ability refers to the capacity to remember the premises, and gist ability refers to the capacity to use the appropriate pattern information to infer the transitive relationship.

Probabilities to retrieve and use verbatim and gist-trace levels depend on verbatim or gist ability levels. The higher the ability level, the higher the probability to retrieve and accurately use the appropriate trace for a specific task and to produce a correct answer. When children have a low ability level, they are expected to retrieve and use irrelevant traces with high probability, and to retrieve and use relevant traces with low probability. When children have high ability level, they retrieve and use irrelevant traces with low probability, and relevant traces with high probability.

Verbatim and gist ability levels are expected to increase with age (Sternberg & Weil, 1980). Brainerd and Kingma (1984, 1985), Reyna and Brainerd (1990), and Reyna (1992) posit that verbatim ability in transitive reasoning develops rather fast and reaches completion at approximately five years of age. Gist ability develops at a slower pace and is not expected to reach full development during childhood. Young school-aged children in particular use verbatim memory while older children rely on gist traces (Reyna & Brainerd, 1990; see also Dayton, 1998; Liben & Posnansky, 1977; Marx, 1985b, 1985a;

Perner & Mansbridge, 1983; Reyna, 1996; Stevenson, 1972). However, age may not completely explain individual differences in verbatim and gist ability but intra-age variability and differences in developmental rates may also have an effect (Wohlwill, 1973). Therefore, studying individual differences and developmental trajectories instead of investigating the abilities using fixed age-groups seems to be mandatory (see also Bouwmeester & Sijtsma, in press).

Suppose a child has high verbatim ability level and low gist ability level and is required to respond to a memory and a transitivity test-pair. Because of his/her high verbatim ability level we expect a high probability of retrieving a verbatim trace containing the correct premise. Because of his/her low gist ability level we expect a low probability of retrieving the correct pattern information. Thus, performance on the memory test-pairs is expected to be good due to the high verbatim ability but performance on transitivity test-pairs is expected to be poor due to the lack of pattern information. Suppose that another child has both low verbatim and gist ability levels. We expect that (s)he has low probability of retrieving a verbatim trace holding the correct premises and also low probability of retrieving a gist trace containing relevant pattern information. Alternatively, (s)he has high probability of using a verbatim trace that contains irrelevant verbatim information (e.g., about the color of the objects when that is irrelevant) which does not lead to correct answers. Suppose that a third child has high gist ability level. This child has high probability of using a gist trace containing relevant pattern information for inferring both the memory and the transitivity test-pair, and this will most likely produce correct answers. Note that here the verbatim ability level is not important: Once a child has a high gist ability level, (s)he is able to solve both the memory and the transitivity test-pairs by using the accurate pattern information.

The summary of our individual-difference model so far is the following. We assume that children differ in their verbatim and gist ability levels due to age differences and intra-age variability. Depending on these ability levels, children are characterized by particular probabilities to retrieve and use particular verbatim and gist-trace levels from a hierarchy of trace levels. The combination of verbatim and gist-trace levels that are retrieved determines performance on memory and transitivity test-pairs.

Retrieval of a particular trace is mediated by task cues, if they are available (Brainerd & Reyna, 1990a). If the ordering of the objects is obvious (e.g., objects are shown simultaneously and ordered in ascending magnitude), a relatively low gist ability level may be sufficient to retrieve an appropriate gist trace which contains all pattern information needed for a correct answer. However, if lack of pattern cues makes it difficult to perceive the ordering of the objects, a relatively low gist ability level may not be sufficient to retrieve the appropriate gist trace. This may result in poor performance on the transitivity test-pairs (Brainerd & Reyna, 1990a). Then it depends on the verbatim ability level whether an accurate verbatim trace can be retrieved for solving the memory test-pairs.

Theoretically, confronted with a particular transitive reasoning task a child processes an unlimited number of possible verbatim and gist-trace levels in parallel, each trace having its own probability to be retrieved given ability level and available task cues. However, only trace levels resulting in different response patterns can be distinguished in empirical data. For example, when a child retrieves verbatim traces containing information about the color of the objects (e.g., "I saw a red, a yellow, a green, an orange, and a purple stick") when in fact length is the property of interest, probabilities of producing correct answers to the memory test-pairs are approximately at chance level. Notice that similar probabilities result

if another child retrieves a different verbatim trace that contains information about the objects' shape (e.g., "the sticks were vertical bars"). The point is that these different verbatim traces each produce the same probability pattern (i.e., probabilities at chance level) and that observed responses do not distinguish the traces that underlay them.

Three verbatim traces and three gist traces are expected to be empirically distinguishable. The first verbatim and gist-trace levels contain irrelevant information resulting in success probabilities approximately at chance level, both for memory and transitivity test-pairs. The third trace levels contain highly relevant information resulting in expected success probabilities close to 1. To formulate the expected success probabilities for the trace levels in between we used Trabasso's work (Trabasso, 1977; Trabasso et al., 1975) and Brainerd's and Kingma's (1984) work. The results of the experiments of Trabasso and his colleagues (Bryant & Trabasso, 1971; Riley, 1976; Riley & Trabasso, 1974; Trabasso & Riley, 1975; Trabasso, 1977; Trabasso et al., 1975) and those of Brainerd and Kingma (1984) both indicated that children produced a particular pattern of errors on the memory and transitivity test-pairs when they did not completely remember the premise information or when they did not completely perceive the pattern of the objects. These particular patterns of errors are used to define the in-between verbatim- and gist-trace levels in the next two subsections. To keep the discussion simple, we describe the verbatim and gist-trace levels independently. That is, when describing verbatim traces we assume gist ability level to be low, and when describing gist traces we assume verbatim ability level to be low. However, notice that in the complete individual-difference model both abilities function simultaneously and that the combination of verbatim and gist trace-levels determines performance on memory and transitivity test-pairs.

Verbatim traces

Fig. 1 shows the relationships between verbatim ability, verbatim traces and performance on memory test-pairs. Note that because the verbatim traces do not contain

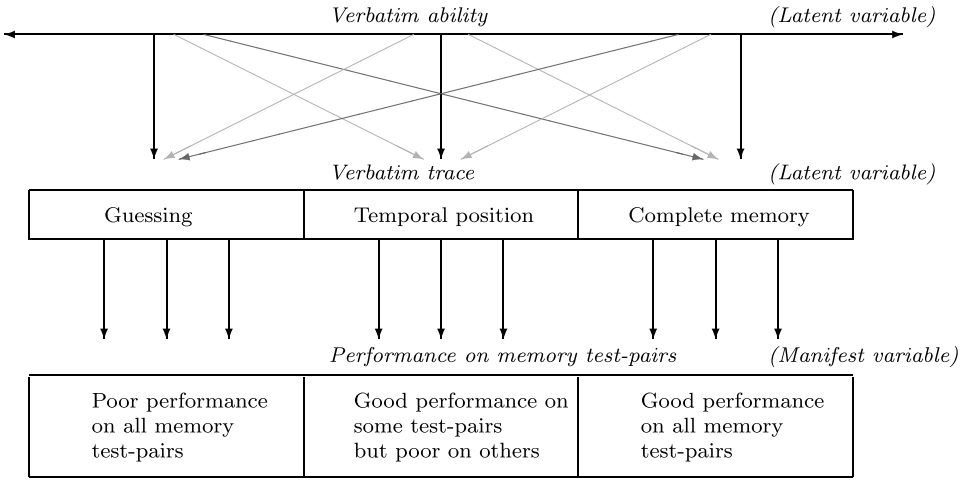


Fig. 1. Relationships between latent verbatim ability, latent verbatim trace, and performance on memory test-pairs.

pattern information, verbatim ability does not influence performance on transitivity test-pairs. The three levels in Fig. 1 are connected by two probability structures: One connects ability levels with trace levels, and the other connects trace levels with memory test-pair performance.

Verbatim ability is hypothesized to induce verbatim traces according to a particular conditional probability structure. The probability distribution is defined as P(verbatim trace | verbatim ability level), which is the probability of using a particular trace given a particular verbatim ability level. Note that both verbatim ability and verbatim traces are unobservable or, in our preferred terminology, latent.

We assume three trace levels. At the first level, the verbatim trace does not contain relevant premises. Instead of such information, for example about the length of sticks, this trace may contain information about the sticks' color or shape. This lack of relevant information has the effect that children guess for the correct answer to each of the memory test-pairs ("guessing" at the second level in Fig. 1) and, as a result, performance probabilities are approximately at chance level.

At the second level, the verbatim trace contains relevant but incomplete premise information. More specifically, based on the work of Trabasso and his collegues (Bryant & Trabasso, 1971; Riley, 1976; Riley & Trabasso, 1974; Trabasso, 1977; Trabasso & Riley, 1975; Trabasso et al., 1975, see also Brainerd & Kingma, 1984) we expect primacy and recency effects to lead to better performance on test pairs for the premises presented first and last than on test pairs for the premises presented in between ("temporal position" in Fig. 1). In children's short-term memory primacy effects tend to be stronger than recency effects (Berch, 1979) and, consequently, this ordering is also expected to occur with the memory test-pairs. Notice that Trabasso and his colleagues used the overlearning paradigm and that their results may not be completely comparable with those for the standard transitive reasoning task. However, although in their experiments performance on the memory test-pairs was almost perfect by the end of the training session, the particular kinds of errors made during training were the same as those made in the standard transitive reasoning task (see e.g., Brainerd & Kingma, 1984).

At the third level, the verbatim trace contains all the relevant premises. This results in high performance probabilities on all memory test-pairs ("complete memory" in Fig. 1).

The conditional probability structure is the following. It is hypothesized that P(guessing trace | ability level) decreases as a function of ability (i.e., with increasing ability it is less likely that the guessing trace is retrieved), and is maximal when ability level is low; P(temporal-position trace | ability level) first increases and then decreases as a function of ability and is maximal when the ability level is intermediate (i.e., the temporal-position trace is characteristic of intermediate ability levels, but rare for low and high levels); and P(complete-memory trace | ability level) increases as a function of ability and is maximal when ability level is high (i.e., the complete-memory trace is most easily retrieved at high ability levels). These hypothesis are visualized in Fig. 1: Solid black arrows between the two latent variable levels indicate high probability, grey dotted arrows indicate lower probability, and light grey dotted arrows indicate low probability.

Gist traces

Fig. 2 shows the relationships between gist ability, gist traces and performance on memory and transitivity test-pairs. The unitary-trace model (Brainerd & Kingma, 1984)

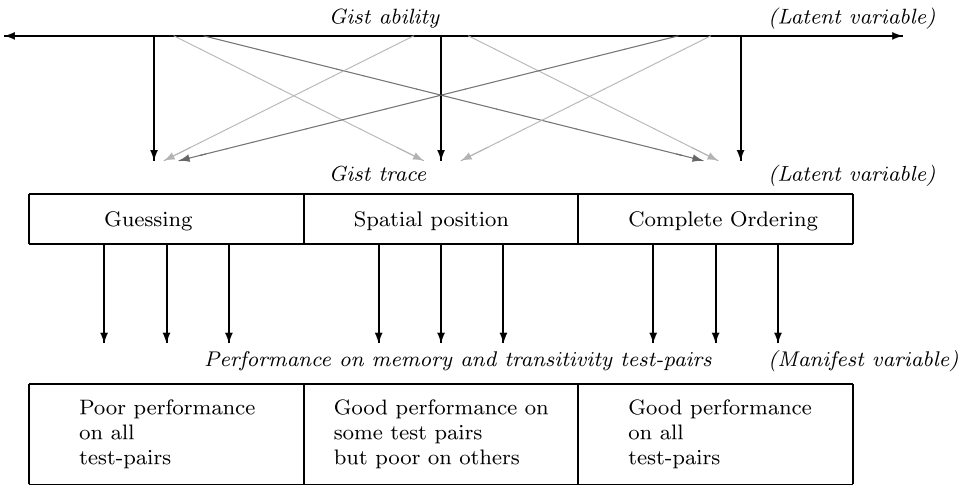


Fig. 2. Relationships between latent gist ability, latent gist trace, and performance on memory and transitivity test-pairs.

predicts that a gist trace affects performance on both memory and transitivity test-pairs. One probability structure connects the ability and trace levels, and another connects the trace and performance levels.

As with verbatim abilities and traces, it is hypothesized that gist ability induces gist traces according to a particular conditional probability structure. Three gist traces were hypothesized. Each gist trace corresponds with a particular probability of answering a memory or a transitivity test-pair correctly.

At the first level, relevant pattern information is absent and children guess for the correct answer on all memory and transitivity test-pairs ("guessing" in Fig. 2). This is likely to result in poor performance.

At the second level, the gist trace contains relevant pattern information but not the complete ordering. Bryant and Trabasso (1971) and Trabasso et al. (1975; see also Brainerd and Kingma, 1984) showed that when forming an internal representation, the end-anchored pairs are learned first followed by the middle pairs. Thus, it is expected that performance on the test pairs in the middle will be poorer than on the end-anchor test pairs. This is a spatial-position effect. Because it uses only part of the ordering information, we expect that this effect will occur at the second trace level ("spatial position" in Fig. 2). The performance probabilities are expected to be high for memory and transitivity test-pairs on both ends of the ordering, and lower for test pairs in between. For example, when five objects are ordered as $Y_A < Y_B < Y_C < Y_D < Y_E$ and a child uses the spatial-position trace, good performance is expected on memory test-pairs $[A, B]$ and $[D, E]$ and transitivity test pairs $[A, C]$ and $[C, E]$; and poorer performance is expected on the other test pairs. The spatial-position effect may also occur at only one end of the ordering; for example, when the gist trace is "right-side objects are large".

At the third level, the trace contains the complete pattern information for reproducing the memory test-pairs and inferring the transitive relationships. The success probabilities on all memory and transitivity test-pairs are expected to be high. The conditional probability structure is the following. It is hypothesized that $P(\text{guessing trace} \mid \text{ability level})$

decreases as a function of ability, and is maximal when ability level is low; P(spatial-position trace | ability level) first increases as a function of ability, reaches its maximum when ability level is intermediate, and then decreases; and P(complete-ordering trace | ability level) increases as a function of ability, and is maximal when ability level is high. These hypothesis are visualized in Fig. 2: Solid black arrows between the two latent variable levels indicate high probability, grey dotted arrows indicate lower probability, and light grey dotted arrows indicate low probability.

Transitive-reasoning tasks and task manipulations

Children may differ in their verbatim and gist ability levels and this will likely result in different performance on transitive reasoning tasks. Transitive reasoning tasks may also vary in difficulty level depending on the complexity of the operations that have to be performed to infer the transitive relationship (Brainerd & Reyna, 1990a). The fewer cues the task offers for perception of pattern information, the more difficult the task. Chapman and Lindenberger (1992) argued that fuzzy trace theory is not valid when tasks offer only few cues or no cues whatsoever for perceiving the ordering of the objects (e.g., in tasks in which the premise pairs are presented successively). Brainerd and Reyna (1992) posit that gist traces can be used to solve tasks in which objects are presented simultaneously and tasks in which objects are presented successively, but that pattern information is easier to perceive in simultaneously presented tasks. Tasks in which the objects are successively presented are more difficult.

Retrieval of a trace containing the appropriate pattern information is easier when the position of the objects is ordered rather than disordered (Verweij, 1994). Also, when presentation of the objects is ordered, trace retrieval is expected to be easier than when presentation is disordered. Reyna and Brainerd (1990) and Brainerd and Reyna (1992) noted that scrambling the premises greatly increases the difficulty of transitivity tasks (see also Chapman & Lindenberger, 1988; Riley, 1976). Note that these task variations are not expected to influence performance on memory test-pairs, because such test pairs can be solved by means of verbatim traces which are not expected to be influenced by variations in pattern cues.

Verbatim and gist traces are processed in parallel. Thus, the combination of each of the three verbatim-trace levels and each of the three gist-trace levels yields nine possible combinations, each of which induces typical performance on memory and transitivity test-pairs. Task characteristics are expected to differentially influence the retrieval of verbatim and gist traces. For example, when objects in a task are *positioned* in a linear order and also *presented* in a linear order the cues on ordering are obvious. As a consequence, the required gist ability level is lower than when, for example, the objects are not positioned or presented in a linear order.

Based on such considerations, we used three kinds of tasks that may be characterized as (1) ordered position, ordered presentation ($O_{\text{pos}}O_{\text{pres}}$); (2) ordered position, disordered presentation ($O_{\text{pos}}D_{\text{pres}}$); and (3) disordered position, ordered presentation ($D_{\text{pos}}O_{\text{pres}}$). The combination of "disordered position, disordered presentation" was not used because it was expected to be too difficult even for adults Brainerd and Reyna, 1992; Verweij, 1994. With each task type, first four premise pairs were presented before children were confronted with four memory test-pairs and three transitivity test-pairs. The description of the task types is as follows.

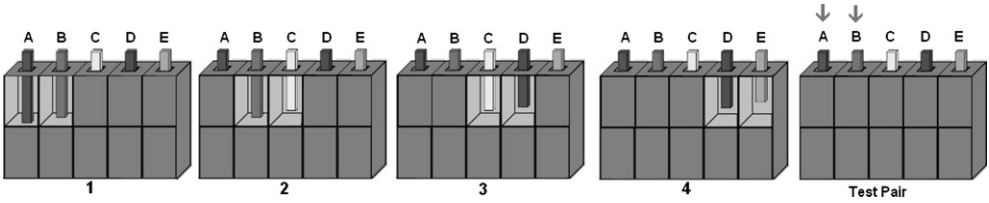


Fig. 3. Example of the premise presentation of an "Ordered Position, Ordered Presentation" task; letters A, B, C, D, and E were not visible to the children.

Ordered position, ordered presentation tasks ($O_{pos}O_{pres}$ tasks)

Objects in $O_{\text{pos}}O_{\text{pres}}$ tasks are ordered from small to large or from large to small. The presentation of the premises is also ordered. Thus, first premise pair $[A, B]$ is presented, followed consecutively by premise pairs $[B, C]$, $[C, D]$, and $[D, E]$. Ordered presentation of ordered objects renders the use of pattern information from gist traces rather easy. Fig. 3 shows the four premises of an $O_{\text{pos}}O_{\text{pres}}$ task. Note that the greys in fact were yellow, green, purple, red, orange, and blue when presented to the children. Box 1 presents the first premise pair, box 2 the second premise pair, and so on. The "test pair" box shows the first memory test-pair $[A, B]$, denoted as M_1.

For combinations of verbatim and gist-trace levels, Table 1 shows the expected performance patterns on the memory and transitivity test-pairs of the $O_{\text{pos}}O_{\text{pres}}$ tasks. When gist-trace level is *intermediate* or *high*, expected performance is good because pattern information can easily be used to infer the relationships in both the memory and transitivity test-pairs. When gist-trace level is *low*, the combination with

- *low* verbatim-trace level is expected to result in guessing, yielding success probabilities at approximately chance level on all memory and transitivity test-pairs;
- *intermediate* verbatim-trace level is expected to produce temporal-position effects, resulting in moderate performance on the first and last memory test-pairs (M_1:[A, B] and M_4:[D, E]) and poor performance on all other memory and transitivity test pairs; and

Table 1
Expected performance on the test pairs of $O_{\text{pos}}O_{\text{pres}}$ tasks for nine combinations of trace levels

Verbatim	Gist	Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○
	Intermediate	•	•	•	•	•	•	•
	High	•	•	•	•	•	•	•
Intermediate	Low	⊗	○	○	⊗	○	○	○
	Intermediate	•	•	•	•	•	•	•
	High	•	•	•	•	•	•	•
High	Low	•	•	•	•	○	○	○
	Intermediate	•	•	•	•	•	•	•
	High	•	•	•	•	•	•	•

○, poor performance; ⊗, moderate performance; •, good performance.

- *high* verbatim-trace level is expected to produce complete memory of the premises, yielding high success probabilities on the memory test-pairs and low success probabilities on the transitivity test-pairs.

Ordered position, disordered presentation tasks ($O_{pos}D_{pres}$ tasks)

In $O_{\text{pos}}D_{\text{pres}}$ tasks, the objects are ordered from small to large or large to small. The presentation of the premise pairs is disordered; for example, in Fig. 4 first $[C, D]$ is presented, followed consecutively by $[A, B]$, $[D, E]$, and $[B, C]$. The midterm relationships are always presented first and last, and the end anchors are always presented in between. Therefore, we are able to distinguish a temporal-position effect from a spatial-position effect in performance on the memory test-pairs (see also Brainerd & Kingma, 1984). Let ○ denote poor performance and ⊗ moderate performance; then for the temporal-position effect, on the four memory test-pairs we expect the pattern (⊗○○⊗) and for the spatial-position effect we expect (○⊗⊗○). Disordered presentation renders the use of gist traces more difficult than ordered presentation because it is more difficult to recognize the ordering of the objects. The "test pair" box in Fig. 4 shows the first transitivity test-pair $[A, C]$, denoted as T_1.

For all nine combinations of verbatim and gist-trace levels, Table 2 shows the expected performance on the memory and transitivity test-pairs of the $O_{\text{pos}}D_{\text{pres}}$ tasks. When verbatim-trace level is *low*, the combination with

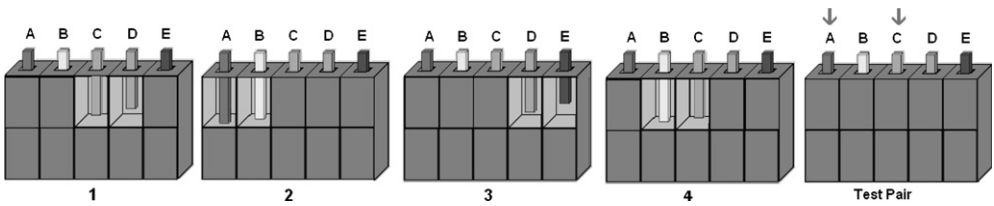


Fig. 4. Example of the premise presentation of an "Ordered Position, Disordered Presentation" task; letters A, B, C, D, and E were not visible to the children.

Table 2
Expected performance on the test pairs of $O_{\text{pos}}D_{\text{pres}}$ tasks for nine combinations of trace levels

Verbatim	Gist	Hypothesized probabilities						
		Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○
	Intermediate	○	⊗	⊗	○	○	⊗	⊗
	High	•	•	•	•	•	•	•
Intermediate	Low	⊗	○	○	⊗	○	○	○
	Intermediate	⊗	⊗	⊗	⊗	○	⊗	⊗
	High	•	•	•	•	•	•	•
High	Low	•	•	•	•	○	○	○
	Intermediate	•	•	•	•	○	⊗	⊗
	High	•	•	•	•	•	•	•

○, poor performance; ⊗, moderate performance; •, good performance.

- *low* gist-trace level results in poor performance on all memory and transitivity test-pairs;
- *intermediate* gist-trace level produces a spatial-position effect which results in moderate performance on the end anchors, M_2:[A, B], M_3:[D, E], T_2:[A, C], and T_3:[C, E]; and poor performance on the other test pairs; and
- *high* gist-trace level produces good performance on all test pairs because the ordering information can be used to solve both memory and transitivity test-pairs.

When verbatim-trace level is *intermediate*, temporal position effects are expected for the memory test-pairs. The combination of intermediate verbatim-trace level with

- *low* gist-trace level produces only temporal-position effects, yielding moderate performance on the first and last presented memory test-pairs (M_1:[B, C] and M_4:[C, D]) and poor performance on all other memory and transitivity test-pairs;
- *intermediate* gist-trace level produces both temporal position and spatial-position effects resulting in moderate performance on all memory and transitivity test-pairs except T_1:[B, D]. That is, for the temporal-position effect we expect the pattern ($\otimes\circ\circ\otimes\ \circ\circ\circ$) (notice that a temporal-position effect does not influence the performance on transitivity test-pairs), and for the spatial-position effect we expect the pattern ($\circ\otimes\otimes\circ\ \circ\otimes\otimes$). When both effects influence performance simultaneously, we expect the pattern: ($\otimes\otimes\otimes\otimes\ \circ\otimes\otimes$); and
- *high* gist-trace level produces good performance on all memory and transitivity test-pairs.

When verbatim-trace level is *high*, the combination with

- *low* gist-trace level produces complete memory of the premises, resulting in high success probabilities. Because pattern information is not available, poor performance is expected on the transitivity test-pairs;
- *intermediate* gist-trace level produces complete memory and a spatial-position effect resulting in good performance on all memory test-pairs and moderate performance on the end-anchored transitivity test-pairs T_2:[A, C] and T_3:[C, E]; and
- *high* gist-trace level leads to good performance on all test pairs.

Disordered position, ordered presentation tasks ($D_{pos}O_{pres}$ tasks)

In $D_{\mathrm{pos}}O_{\mathrm{pres}}$ tasks, the objects are positioned disorderly. For example, in Fig. 5 stick A is in the third position in the box and stick B is in the first position. The presentation of the premises is ordered. In Fig. 5, first premise pair $[A, B]$ is presented, followed consecutively by premise pairs $[B, C]$, $[C, D]$, and $[D, E]$. Because positional cues about the ordering of the objects are not provided, a disordered position is expected to require both high verbatim and gist-ability levels. Consequently, not only the ordering has to be recognized but also the premises have to be memorized. The "test pair" box (Fig. 5) shows the first memory test-pair.

Table 3 shows the expected performance patterns on the memory and transitivity test-pairs of $D_{\mathrm{pos}}O_{\mathrm{pres}}$ tasks for all combinations of verbatim and gist-trace levels. When verbatim-trace level is *low* (i.e., when the guessing trace operates), performance is expected to be poor on all memory and transitivity test-pairs independent of gist-trace level. At least

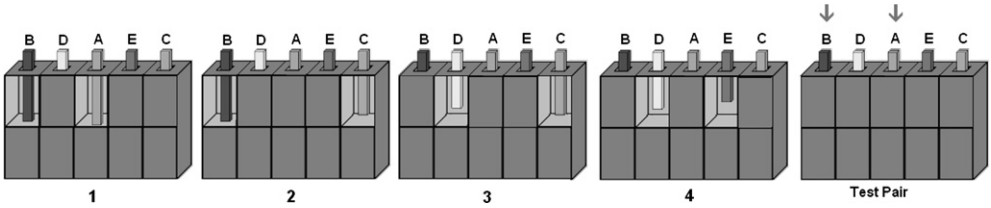


Fig. 5. Example of the premise presentation of a "Disordered Position, Ordered Presentation" task; letters A, B, C, D, and E were not visible to the children.

Table 3
Expected performance on the test pairs of $D_{pos}O_{pres}$ tasks for nine combinations of trace levels

Verbatim	Gist	Hypothesized probabilities						
		Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○
	Intermediate	○	○	○	○	○	○	○
	High	○	○	○	○	○	○	○
Intermediate	Low	⊗	○	○	⊗	○	○	○
	Intermediate	⊗	○	○	⊗	○	○	○
	High	⊗	○	○	⊗	⊗	○	⊗
High	Low	•	•	•	•	○	○	○
	Intermediate	•	•	•	•	○	○	○
	High	•	•	•	•	•	•	•

○, poor performance; ⊗, moderate performance; •, good performance.

intermediate verbatim ability level is needed to remember the premises or recognize the ordering of the objects. When verbatim-trace level is *intermediate*, the combination with

- *low* and *intermediate* gist-trace levels produces temporal-position effects resulting in moderate performance on the first and last presented memory test-pairs (M_1:[A, B] and M_4:[D, E]); and
- *high* gist-trace level produces spatial-position effects yielding moderate performance on the end-anchors (M_1:[A, B], M_4:[D, E], T_1:[A, C], and T_3:[C, E]).

When verbatim-trace level is *high*, the combination with

- *low* and *intermediate* gist-trace levels produces complete memory resulting in good performance on the memory test-pairs but poor performance on transitivity test-pairs; and
- *high* gist-trace level produces good performance on all memory and transitivity test-pairs.

Theoretical model and research questions

Fig. 6 shows the individual-difference model of fuzzy trace theory with respect to transitive reasoning. At the highest level (technically referred to as the third level of analysis)

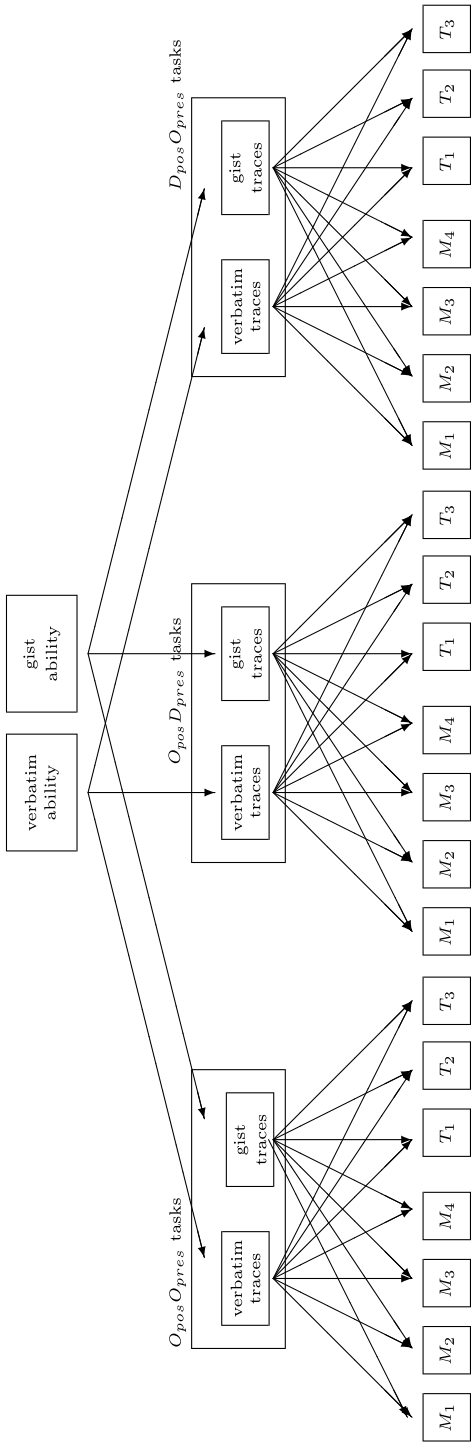


Fig. 6. Individual-difference model of fuzzy trace theory for transitive reasoning.

are the verbatim and gist-ability levels. These abilities govern the verbatim and gist traces through conditional probability processes. The traces constitute the second level of analysis. The abilities and the traces are latent (i.e., unobservable). The traces govern the probabilities of manifest (i.e., observable) correct and incorrect responses to the memory and transitivity test-pairs of the three kinds of tasks. This is the first level of analysis.

Abilities are usually assumed to be continuous (Embretson & Reise, 2000). For transitive reasoning we have no reason to deviate from this general assumption; thus, statistically the verbatim and gist abilities are considered continuous latent variables. Further, we have distinguished three ordered verbatim traces and three ordered gist traces. Statistically, these triplets of latent traces can be conceived of as ordered categorical latent variables.

The individual-difference model for transitive reasoning (Fig. 6) fits our data when the estimated probabilities of performance on test pairs are similar in relative magnitude to the hypothesized probabilities (see Tables 1–3). A fitting model explains individual differences in performance on memory and transitivity test-pairs due to the differential use of verbatim and gist-trace levels.

Brainerd and Kingma (1984) provided a sophisticated description of fuzzy trace theory and its observable consequences for different task manipulations, but the statistical methods available at the time did not allow modeling of response patterns and individual differences. They tested fuzzy trace theory on a priori identified fixed age groups, and used mean scores per age group to test hypotheses about verbatim and gist traces. Thus, they were able to test the contours of fuzzy trace theory but their methodology did not enable them to distinguish individual differences or groups of individuals using the same combination of verbatim and gist traces. In this study, we used modern latent class analysis (Vermunt, 2003) to distinguish groups of children that differ in their use of verbatim and gist traces when responding to memory and transitivity test-pairs. The data for the latent class analysis were the patterns of incorrect/correct scores (score 0 for an incorrect answer and score 1 for a correct answer) on the four memory test-pairs and the three transitivity test-pairs in each task; there were 12 tasks in total, to be discussed shortly.

First, the fit of the complete fuzzy trace model (as formulated in Fig. 6) to these data was investigated. The fit of the model was compared to that of several competing models which reflected alternative data structures derived from Piaget's theory (Piaget, 1942, 1947) and linear ordering theory (Trabasso et al., 1975). Second, at a more detailed level the performance on each of the test pairs as predicted by the fuzzy trace model was compared to the performance reflected by the empirical data. Third, the relationship between age and verbatim and gist ability was investigated.

From theoretical model to statistical model

Multilevel latent class modeling (Vermunt, 2003) was used to evaluate the fit of the theoretical model (Fig. 6) to the data. This method was preferred over the traditional and much-used analysis of variance (ANOVA) model for three reasons. First, the manifest dependent variables are binary incorrect/correct scores, whereas ANOVA assumes dependent variables to be continuous and normally distributed. This assumption is unrealistic in our application. Second, at two levels in the theoretical model the data are dependent or nested within the levels of higher-order variables: (1) the

seven test-pair scores within a task—four memory test-pair scores and three transitivity test-pair scores—are mutually dependent to some degree due to the combinations of trace levels that are retrieved. For example, when a child retrieves the gist trace "objects become smaller from right to left", (s)he is able to infer all memory and transitivity test-pairs correctly; and (2) the combination of verbatim and gist traces (intermediate level) that is used for solving a particular task is dependent on the child's verbatim and gist ability levels (highest level). A multilevel model incorporates these dependencies, whereas a within-subject ANOVA is unable to do this. Ignoring this dependence reduces the power of statistical tests. Third, the theoretical model encompasses both manifest and latent variables, whereas an ANOVA model can deal only with manifest variables. To summarize, multilevel latent class models are more appropriate for evaluating the fit of the theoretical model to the data than ANOVA models.

The basis of the multilevel latent class model is the simple latent class model (Clogg, 1988; Heinen, 1996; McCutcheon, 1987; see also Jansen & Van der Maas, 1997; Raijmakers, Jansen, & Van der Maas, 2004; Rindskopf, 1987). A latent class model may be interpreted as a factor analysis model for categorical data, such as the 0/1 scores on the test pairs, which usually results in a limited number of homogeneous subgroups of respondents. Because these subgroups are identified from the patterns of incorrect/correct scores rather than from an a priori defined sub-grouping based on age or educational level, they are called latent classes.

Latent class analysis of 0/1 scores from the memory and transitivity test-pairs yields two kinds of estimated probabilities. First, there are so-called class probabilities which add up to 1 and reflect the relative size of each of the latent classes. Suppose a latent class model with three classes is found to fit the data best, and these classes have class probabilities equal to .35, .40 and .25. Applied to our study, these probabilities would mean, for example, that on a particular task 35% of the children retrieve a low verbatim-trace level (i.e., "guessing trace"), 40% retrieve an intermediate verbatim-trace level (i.e., "temporal-position trace"), and 25% retrieve a high verbatim-trace level (i.e., "complete-memory trace").

Second, consider the three transitivity test-pairs for each task in this study. For each transitivity test-pair, each latent class is characterized by a conditional probability of a correct answer given membership in that class. For example, within a particular latent class the conditional probabilities of producing correct responses to the transitivity test-pairs $[A, C]$, $[B, D]$, and $[C, E]$ of the $O_{\text{pres}}O_{\text{pos}}$ tasks are found to be .9, .5, and .9, respectively. The members of this latent class—here assumed to correspond to a particular trace level—will respond correctly with high probability to the first and last transitivity test-pairs while their responses to the middle transitivity test-pairs can go either way. These conditional response probabilities can be used to understand performance typical in this latent class. For example, the (artificial) conditional probabilities given here may suggest a spatial-position trace. Conditional probabilities are unlikely to reach the value of 1 because some children may become bored, tired, or lose concentration, even if they have high ability levels.

Conditional response probabilities help to interpret the latent classes. Their role is comparable to the role patterns of factor loadings play in establishing the meaning of factors in a factor analysis. The class probabilities and the conditional response probabilities together can be used to determine for each respondent in which latent class (s)he belongs most

likely on the basis of her/his pattern of 0/1 scores on the memory and transitivity test-pairs.

Our statistical model is an extension of the simple latent class model (see also Raijmakers et al., 2004) in two respects. First, the latent class model for the responses to a particular task is connected to the latent abilities by means of a multilevel structure; hence, a multilevel latent class model is used. Second, the probability of retrieving a particular discrete trace level is modeled as a function of the continuous ability level. This is done by means of the partial credit model (Masters, 1982); this is an item response model that is particularly suited for explaining an ordered, discrete dependent variable (here, a trace) from an independent continuous variable (here, an ability). Appendix A provides a technical description of the multilevel latent class model, including the two extensions (i.e., the multilevel model and the partial credit model).

The program Latent Gold (Vermunt & Magidson, 2005) was used to estimate the class probabilities and the conditional response probabilities of the model, and compute the fit statistics. For evaluating the fit of the model to the data, the sample was randomly split into two halves. The first half was used to evaluate the improvement of the fit of different, competing models. Next, the fit of the hypothesized model estimated in the first half of the sample was compared with the fit of the same model in the second half. When the fit statistics in both halves were close, the degree of chance capitalization was small and the model could be considered valid for the population.

General data analysis procedure

The data analysis is divided and discussed in three sections. First, the fuzzy trace model was estimated from the data and the fit of the model to the data was compared with that of alternative models for transitive reasoning. Second, the estimated probability structure was compared with the probability structure expected from the theory (see Tables 1–3). Third, the relationship between age and estimated ability level was investigated. Before that, descriptions are given of the instrument, the sample, the procedure, and the design of the study, and some preliminary analyses are discussed.

Method

Instrument

An individual computer test for transitive reasoning was constructed (Bouwmeester & Aalbers, 2004). Binary performance scores were registered automatically during test administration. Four test versions each presented the tasks in a different order. Each child was administered one randomly chosen version. The use of four versions was meant to rule out order effects due to task presentation. This was checked statistically by means of an ANOVA.

Sample

The transitive reasoning test was administered to 409 children ranging from 5 to 13 years of age. Children came from four elementary schools in the Netherlands. They were from middle class social-economic status families. Table 4 shows the number of

Table 4
Number of children, mean age in months (*M*) and standard deviation (*SD*) in each grade

Grade	Number	Age	
		M	*SD*
Kindergarten	39	73.67	4.70
1	65	86.15	4.81
2	70	100.16	5.85
3	60	111.80	5.80
4	63	123.44	5.52
5	58	140.31	7.69
6	54	146.18	6.61

children per grade, and the mean age and the standard deviation of age within each grade.

Design

Three types of tasks were used: $O_{pos}O_{pres}$, $O_{pos}D_{pres}$, and $D_{pos}O_{pres}$. Four versions of each task type were administered; thus, there were 12 tasks in total. The four tasks of the same type differed with respect to the colors of the sticks, and with respect to the direction of the ordering or the presentation; that is, sticks could be ordered from left to right or from right to left, and they could be presented from small to large, or from large to small. One task type was always followed by a different type. We did not use a fully randomized design because the tasks differed clearly in difficulty level. For example, the $O_{pos}O_{pres}$ tasks were much easier than the $D_{pos}O_{pres}$ tasks, and in a fully randomized design it would have been possible that children would be consecutively administered three difficult $D_{pos}O_{pres}$ tasks. We expected that this would discourage especially the younger children. Based on previous research (Bouwmeester & Sijtsma, in press), we expected that the bias in the results caused by lack of motivation would be much greater than possible bias caused by non-randomization.

The presentation of the premises was followed by the presentation of the four memory test-pairs and the three transitivity test-pairs, respectively. The ordering of the memory test-pairs was always the same as the ordering in which the premises had been presented. A 1-score was assigned when the child touched the correct stick on the pc-screen; and a 0-score was assigned otherwise. For each child, 7 (test pairs) × 3 (task types) × 4 (task-type versions) = 84 scores were collected.

Procedure

The test was administered in a quiet room in the school building. The experimenter started a short conversation with the child to put her/him at ease. Two introductory tasks were presented in which it was explained that each time the child had to touch the longest stick. Next, the experimenter explained that there were 13 additional tasks and that the child had to try these tasks on her/his own. The child did not know that the first of the 13 tasks was another introductory task that was meant to let her/him get used to the idea that (s)he had to work on her/his own now.

Table 5
Means (*M*, aggregated over test pairs of the same task), standard errors (*SE*) and 95% confidence intervals (CI) for the task versions (Denoted as A, B, C and D) of each of the three task types

Version	$O_{pos}O_{pres}$			$O_{pos}D_{pres}$			$D_{pos}O_{pres}$		
	M	*SE*	95% CI[a]	*M*	*SE*	95% CI[a]	*M*	*SE*	95% CI[a]
A	.75	.02	.71–.78	.72	.01	.68–.75	.57	.01	.54–.60
B	.78	.01	.74–.81	.69	.01	.66–.72	.59	.01	.56–.62
C	.79	.01	.76–.82	.65	.01	.62–.69	.55	.01	.52–.58
D	.78	.01	.75–.82	.75	.01	.72–.79	.57	.01	.54–.60

[a] Bonferroni adjustment.

Preliminary analysis

An ANOVA was performed to check for possible order effects of the tasks on the number-correct score for the 84 items in total. Number-correct score served as dependent variable and test-version as independent variable. It was found that the four test versions did not differ significantly [$F(3, 401) = 1.32$, $p > .05$]. Thus, presentation order of tasks had no effect on number-correct score.

A within-subject ANOVA was done to test whether the four task versions of the three task types differed with respect to number-correct score. Table 5 shows the means (aggregated over test pairs) and the 95% confidence intervals. For $O_{pos}O_{pres}$ tasks, task versions differed significantly [$F(2.75, 1112.58) = 2.93$, $p = .037$] but partial η^2 (for effect size; Cohen, 1977) was small (.007) and confidence intervals overlapped. For $O_{pos}D_{pres}$ tasks, task versions differed significantly [$F(2.64, 1069.67) = 15.60$, $p = .000$] but partial η^2 was small (.037) and confidence intervals overlapped. For $D_{pos}O_{pres}$ tasks, task versions differed significantly [$F(2.98, 1202.23) = 3.46$, $p = .016$] but partial η^2 was small (.007) and confidence intervals overlapped. Although a few task versions differed significantly with respect to average performance, effect sizes were small and confidence intervals showed that in all cases differences between task versions were small.[1] It was concluded that all task versions could be used to estimate the model.

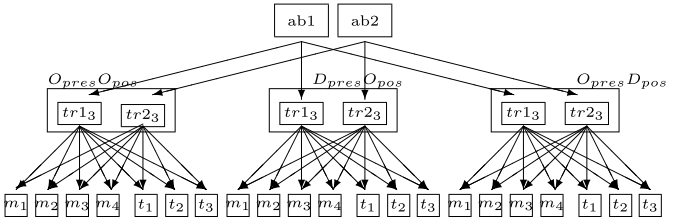
Results

Analysis section 1: Fitting structural models

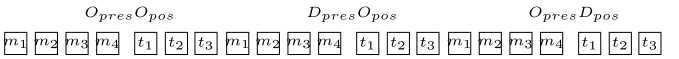
Alternative models for fuzzy trace theory

The individual-difference model of fuzzy trace theory is represented as Model A in Fig. 7. Other models were the following. Fuzzy trace theory assumes continuous verbatim and gist abilities, and discrete verbatim and gist traces each with three ordered levels. Model B is much simpler in that it lacks a latent variable structure. This model resembles an ANOVA on average scores for the three task types, and agrees with how Brainerd and Kingma (1984, 1985) tested their hypotheses; that is, the analysis of average scores allows

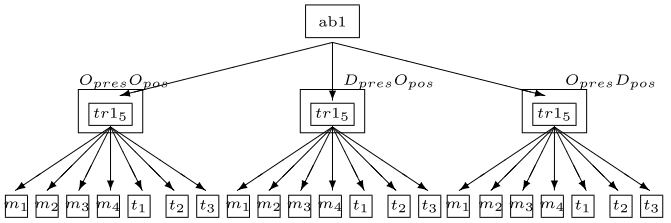
[1] Note that a significant overall *F*-value does not guarantee that individual groups differ significantly (Stevens, 1996, pp. 163–164).



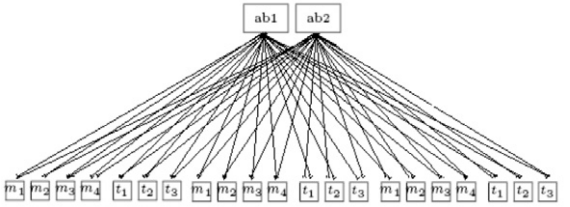
Model A: Individual-difference model of fuzzy trace theory.



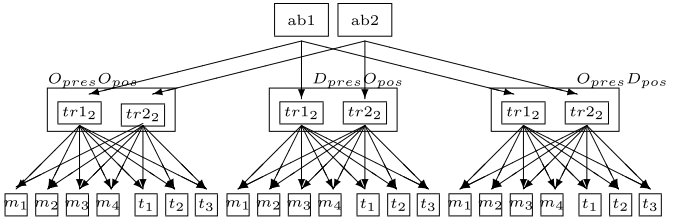
Model B: Model having only observable variables.



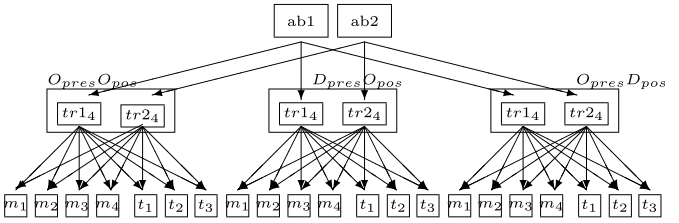
Model C: Model having one ability.



Model D: Model without trace levels.



Model E: Special case of Model A with 2 levels per trace.



Model F: Extension of Model A with 4 ordered levels per trace.

Fig. 7. Six models that were fitted to the data. Explanation of notation: ab$k$, ability k; trl_r, trace l, with r levels; m_s, memory test-pair s; t_f, transitivity test-pair f.

global hypotheses to be tested but not to distinguish different performance groups by means of their responses to different test pairs.

Linear ordering theory predicts that one ability suffices to form a complete linear ordering. Experience with latent class analysis has shown that fitting more than five ordered classes does not improve the fit of the model anymore (Van Onna, 2002). Thus, linear ordering theory with five latent classes is represented by Model C.

Piaget's theory predicts that because each task involves the use of the same rules of logic, task manipulations do not influence performance on test pairs. This is represented by Model D, which assumes a direct effect of ability level on performance without intervening trace levels.

Our individual-difference model (i.e., Model A) assumes that three verbatim-trace levels and three gist-trace levels are optimal. This model is tested against models having either two trace levels (Model E) or four trace levels (Model F).

Specific results

Table 6 shows the fit results for models A through F. Model B, that consist only of manifest variables for task type and item type, fitted poorer than model C that had one ability governing five ordered trace levels [Table 6; see fit statistics *BIC*, *AIC*3, and the decrease in log-likelihood (denoted *LL*)]. The decrease in *LL* given the increase in number of parameters of Model A relative to Model C was substantial; thus, Model A fitted better than Model C. Model D, that formalizes a direct effect of ability level on performance, fitted poorer than model A (see *BIC*, *AIC*3, and decrease in *LL*). Thus, the trace levels cannot be omitted as model D suggests. Model A (three trace levels) fitted better than model E (two trace levels) but model F (four trace levels) fitted better than model A. However, two of the four classes in Model F had nearly the same interpretation; thus, three trace levels seemed to be optimal to distinguish relevant groups. We conclude that Model A describes the data best.

Chance capitalization was evaluated by fitting Model A to the second random half of the sample (Table 6). Because due to missing values the numbers of records (subjects × items per subject) were not exactly the same in both subsamples (2426 and 2434 records), we compared the mean *LL* per record. For the first sample, the mean *LL* equalled −3.66, and for the second sample it equalled −3.73. Thus, Model A fitted almost equally well in both samples.

Table 6
Fit measures for the estimated models

Model	Description	*LL*	*#Par*	*BIC*	*AIC*3
A	2 abilities, 2 traces containing 3 levels	−8880.33	69	18298.45	17967.67
B	Full independence between observed responses	−35633.15	8	71341.27	71290.30
C	1 ability, 1 trace containing 5 levels	−9014.70	47	18395.71	18164.49
D	2 abilities, without mediation by traces	−9259.08	63	19009.18	18707.18
E	2 abilities, 2 traces containing 2 levels	−8914.07	67	18350.34	18029.14
F	2 abilities, 2 traces containing 4 levels	−8848.05	71	18249.47	17909.10
Cross	Model A	−9090.06	69		

LL, log Likelihood (goodness of fit statistic); *#Par*, number of parameters in the model.
BIC, $-2LL + \#par \times \ln(N)$ (for $N = 204$); *AIC*3: $-2LL + 3 \times \#par$.
Both *BIC* and *AIC*3 are information criteria for comparing alternative models.

Discussion

The result of the first analysis section showed that the fit of model A—the individual-difference fuzzy trace theory model—was better than the fit of alternative models B, C, D, and E. It may be noted that Piaget's theory and linear ordering theory are not as explicit with respect to transitive reasoning as fuzzy trace theory; thus, the formalized models for these alternative theories are more liable to subjective interpretation than the fuzzy trace theory model. Nevertheless, the aspects of these alternative theories of which we were certain were represented in the formal models and if they had been valid, they should have been reflected in the estimated data structure.

Analysis section 2: Reproduction of probability structure

For all test pairs, the estimated performance probabilities on the memory and transitivity test-pairs of model A were compared with the expected performance probabilities (Tables 1–3).

Specific results

Global comparison of expected and estimated probabilities. Table 7 shows the structure of the estimated performance probabilities for the seven test pairs in each task, for each combination of verbatim and gist-trace levels. Standard errors of the estimated probabilities (not tabulated here) were between 0.000 and 0.077 (mean = 0.03, standard deviation = 0.02); thus, estimation was accurate.

For a convenient presentation of multiple results, estimated probabilities were summarized in three categories: Notation ○ means probability is lower than .65; ⊗ means probability is between .65 and .80; and • means probability is higher than .80.

For the memory test-pairs (M_1, M_2, M_3, and M_4) a low gist-trace level in combination with (1) a low verbatim-trace level produced low probabilities (rows 1, 2, 3); (2) an intermediate verbatim-trace level produced higher probabilities (rows 10, 11, 12); and (3) a high verbatim-ability level produced high probabilities (rows 19, 20, 21). This pattern was not found for the transitivity test-pairs (T_1, T_2, and T_3). Because ability influenced performance on memory test-pairs but not on transitivity test-pairs, the first latent ability in Model A could be interpreted as verbatim ability.

For low gist-trace level (Table 7, second column), probabilities for transitivity test-pairs (T_1, T_2, and T_3) are low (rows 1, 2, 3; 10, 11, 12; and 19, 20, 21); for intermediate gist-trace level in general probabilities are higher (rows 4, 5, 6; 13, 14, 15; and 22, 23, 24); and for high gist-trace level in general probabilities are highest (rows 7, 8, 9; 16, 17, 18; and 25, 26, 27). Because ability influenced both performance on memory and transitivity test-pairs, the second ability in Model A could be interpreted as gist ability.

Fig. 8a shows the distribution of verbatim-trace levels given verbatim ability. As expected, the probability of using a low verbatim-trace level is maximal when verbatim-ability level is low and decreases as ability increases; the probability of using an intermediate verbatim-trace level first increases and then decreases as a function of ability and is maximal when verbatim-ability level is intermediate; and the probability of using a high verbatim-trace level increases as a function of ability and is maximal when ability level is high. Fig. 8b shows the distribution of gist-trace levels given gist ability. The interpretation is the same as that for the verbatim-trace levels.

Table 7
Global estimated success probability for the test pairs of three task-types, for nine combinations of latent trace levels

Verbatim trace	Gist trace	Task type	Estimated probabilities						
			Memory				Transitivity		
			M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	$O_{pos}O_{pres}$	○	○	○	○	○	○	○
		$O_{pos}D_{pres}$	○	○	○	○	○	○	○
		$D_{pos}O_{pres}$	○	○	○	○	○	○	○
	Intermediate	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	⊗	⊗	•	•	⊗	⊗	•
		$D_{pos}O_{pres}$	○	○	○	○	⊗	○	○
	High	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	•	•	•	•	•	•	•
		$D_{pos}O_{pres}$	⊗	○	○	○	⊗	○	⊗
Intermediate	Low	$O_{pos}O_{pres}$	•	•	•	•	○	○	○
		$O_{pos}D_{pres}$	•	•	•	⊗	○	○	○
		$D_{pos}O_{pres}$	•	⊗	○	○	○	○	○
	Intermediate	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	•	•	•	•	⊗	•	•
		$D_{pos}O_{pres}$	•	•	⊗	⊗	⊗	○	⊗
	High	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	•	•	•	•	•	•	•
		$D_{pos}O_{pres}$	•	•	⊗	•	•	○	⊗
High	Low	$O_{pos}O_{pres}$	•	•	•	•	○	○	○
		$O_{pos}D_{pres}$	•	•	•	•	○	○	○
		$D_{pos}O_{pres}$	•	•	•	•	⊗	○	○
	Intermediate	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	•	•	•	•	⊗	•	•
		$D_{pos}O_{pres}$	•	•	•	•	•	⊗	⊗
	High	$O_{pos}O_{pres}$	•	•	•	•	•	•	•
		$O_{pos}D_{pres}$	•	•	•	•	•	•	•
		$D_{pos}O_{pres}$	•	•	•	•	•	⊗	•

○, <.65; ⊗, .65–.79; •, >.79.

Detailed comparison of expected and estimated probabilities per task. Table 8 shows the hypothesized and the estimated performance probabilities of the test pairs of the four $O_{pos}O_{pres}$ tasks. The majority of the estimated probability patterns agreed with the hypothesized patterns. However, in the fourth row the patterns of the hypothesized and the estimated probabilities differed for the memory test-pairs: It was hypothesized that *intermediate* verbatim-trace level and *low* gist-trace level produce a temporal-position effect, thus predicting moderate probabilities for the memory test-pairs presented first and last (i.e., M_1 and M_4) and low probabilities for the test pairs in between (i.e., M_2 and M_3). However, the high probabilities found suggest complete memory of the premises.

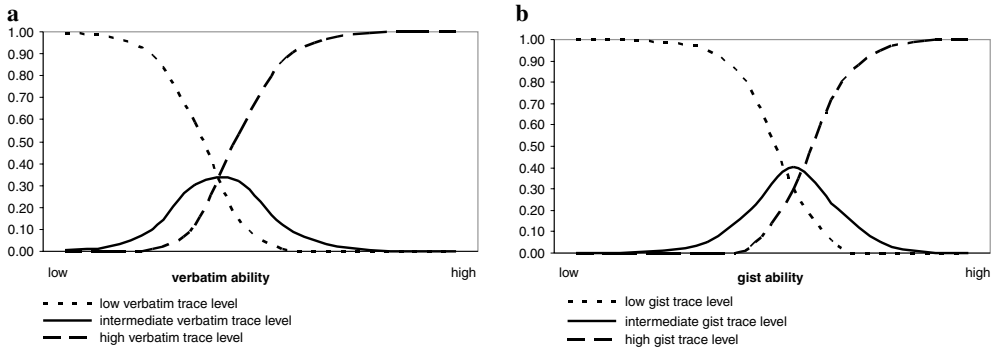


Fig. 8. Distribution of verbatim-trace levels given verbatim ability (a) and distribution of gist-trace levels given gist ability (b).

Table 8
Estimated success probability for the test pairs of tasks $O_{pos}O_{pres}$ for nine combinations of latent trace levels

Verbatim	Gist	Hypothesized probabilities							Estimated probabilities						
		Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○	.59	.48	.38	.49	.55	.56	.50
	Intermediate	•	•	•	•	•	•	•	.94	.97	.95	.96	.97	.94	.97
	High	•	•	•	•	•	•	•	.99	1.00	1.00	1.00	1.00	.99	1.00
Intermediate	Low	⊗	○	○	⊗	○	○	○	.93	.95	1.00	.99	.49	.46	.48
	Intermediate	•	•	•	•	•	•	•	.99	1.00	1.00	1.00	.97	.91	.96
	High	•	•	•	•	•	•	•	1.00	1.00	1.00	1.00	1.00	.99	1.00
High	Low	•	•	•	•	○	○	○	.99	1.00	1.00	1.00	.43	.36	.46
	Intermediate	•	•	•	•	•	•	•	1.00	1.00	1.00	1.00	.96	.87	.96
	High	•	•	•	•	•	•	•	1.00	1.00	1.00	1.00	1.00	.99	1.00

○, <.65; ⊗, .65–.79; •, >.79.

Table 9 shows that for the four $O_{pos}D_{pres}$ tasks the majority of the estimated probability patterns agreed with the hypothesized patterns but that patterns differed in rows 2 and 4. It was hypothesized that *low* verbatim-trace level and *intermediate* gist-trace level together (row 2) produce a spatial-position effect which results in higher probabilities for the end-anchored test-pairs than for the mid-term test-pairs (Table 9, row 2; the end-anchored test-pairs were M_2, M_3, T_2 and T_3). However, the estimated probabilities showed that this spatial-position effect was only active on one end-anchor, which led to high probabilities for the test pairs M_3 and T_3. Also, it was hypothesized that *intermediate* verbatim-trace level and *low* gist-trace level together (row 4) produce a temporal-position effect, but the estimated probabilities showed such an effect only for the first memory test-pairs (M_1 and M_2).

For the four $D_{pos}O_{pres}$ tasks, Table 10 shows that four estimated probability patterns agreed with the hypothesized patterns (in rows 1, 2, 7, and 9). Five patterns (in rows 3, 4, 5, 6, 8) were different. First, for *low* verbatim-trace level and *high* gist-trace level (Table 10, row 3), for all test pairs low probabilities were hypothesized. However, the estimated

Table 9
Estimated success probability for the test pairs of tasks $O_{pos}D_{pres}$ for nine combinations of latent trace levels

Verbatim	Gist	Hypothesized probabilities							Estimated probabilities						
		Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○	.45	.41	.50	.46	.47	.44	.46
	Intermediate	○	⊗	⊗	○	○	⊗	⊗	.71	.74	.87	.82	.79	.79	.89
	High	•	•	•	•	•	•	•	.88	.92	.98	.96	.94	.95	.99
Intermediate	Low	⊗	○	○	⊗	○	○	○	.95	.85	.80	.73	.45	.54	.55
	Intermediate	⊗	⊗	⊗	⊗	○	⊗	⊗	.98	.96	.97	.94	.78	.85	.92
	High	•	•	•	•	•	•	•	.99	.99	1.00	.99	.94	.96	.99
High	Low	•	•	•	•	○	○	○	1.00	.98	.94	.89	.43	.63	.64
	Intermediate	•	•	•	•	○	⊗	⊗	1.00	.99	.99	.98	.77	.89	.94
	High	•	•	•	•	•	•	•	1.00	1.00	1.00	1.00	.93	.97	.99

○, <.65; ⊗, .65–.79; •, >.79.

Table 10
Estimated success probability for the test pairs of tasks $D_{pos}O_{pres}$ for nine combinations of latent trace levels

Verbatim	Gist	Hypothesized probabilities							Estimated probabilities						
		Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
Low	Low	○	○	○	○	○	○	○	.50	.24	.25	.19	.50	.57	.47
	Intermediate	○	○	○	○	○	○	○	.61	.38	.32	.39	.66	.56	.62
	High	○	○	○	○	○	○	○	.71	.54	.41	.64	.79	.56	.75
Intermediate	Low	⊗	○	○	⊗	○	○	○	.88	.71	.63	.52	.61	.63	.53
	Intermediate	⊗	○	○	⊗	○	○	○	.92	.83	.71	.75	.75	.63	.67
	High	⊗	○	○	⊗	⊗	○	⊗	.94	.90	.78	.89	.85	.63	.78
High	Low	•	•	•	•	○	○	○	.98	.95	.90	.84	.71	.69	.58
	Intermediate	•	•	•	•	○	○	○	.99	.97	.93	.93	.82	.69	.71
	High	•	•	•	•	•	•	•	.99	.99	.95	.98	.90	.69	.82

○, <.65; ⊗, .65–.79; •, >.79.

probabilities suggested a spatial-position effect which resulted in moderate and high success probabilities for the test pairs M_1, M_4, T_1, and T_3.

Second, for *intermediate* verbatim-trace level and *low* or *intermediate* gist-trace levels (Table 10, rows 4 and 5) temporal-position effects were hypothesized. However, for *low* gist-trace level the estimated probabilities only showed this effect on the first memory test-pair but not on the last one. For *intermediate* gist-trace level a spatial-position effect was active (in particular on the side where the sticks were longest).

Third, for *intermediate* verbatim-trace level and *high* gist-trace level a spatial-position effect was hypothesized (Table 10, row 6). The estimated probabilities showed a spatial-position effect only on one of the two end-anchors.

Finally, it was hypothesized that *high* verbatim-trace level and *intermediate* gist-trace level produce high memory test-pair probabilities and low transitivity test-pair probabili-

ties (Table 10, row 8). The estimated probabilities for the transitivity test-pairs were high for the end-anchors thus indicating a spatial-position effect.

Discussion

Based on the estimated probability structures, the two abilities could be interpreted as verbatim and gist abilities. At a detailed level observed performance agreed well with hypothesized performance for $O_{pos}O_{pres}$ and $O_{pos}D_{pres}$ tasks but not as well for $D_{pos}O_{pres}$ tasks.

Some relevant deviations from the expected probability patterns are the following. Temporal-position effects were found only for the premises presented first instead of both premises presented first and last. This result agreed with that of Berch (1979), who showed that children's short-term memory primacy effects usually are stronger than the recency effects. Spatial-position effects in $D_{pos}O_{pres}$ tasks were found in particular on one side of the ordering (containing the longest sticks) but not on both sides. This could be due to a marking effect; that is, linguistic factors may have played a role in the end-anchoring. During the premise presentation children had to touch the longest stick on the pc-screen, and this may explain why their representation of the long end-anchor is better than that of the short end-anchor (see Riley & Trabasso, 1974; Sternberg, 1980; Trabasso et al., 1975).

An important result was that children with a high verbatim-ability level but a low gist-ability level performed well on the memory test-pairs but only at chance level on the transitivity test-pairs. This finding disagrees with Trabasso's linear ordering theory which assumes that memory of the premises is sufficient to infer the transitive relationship. Thus, we found that a high ability to remember premises is not enough to correctly infer a transitive relationship. These results agreed with the results found by Halford and Galloway (1977, 1979) and Halford (1979).

Piaget's theory assumes that memory of the premises is a prerequisite for using rules of logic and inferring transitive relationships. Task format was not expected to influence the use of rules of logic when the premises could be remembered. However, we found that memory of the premises was not a prerequisite for inferring the transitive relationship and that task type had a strong influence on the probability of inferring the transitive relationships correctly, even when the premises were remembered correctly. These results contradict Piaget's theory. It may be noted that Piaget's initial aim was not to give a detailed description of transitive reasoning; this renders a comparison between his theory and the present study somewhat disputable.

The results showed that model A fitted poorest for task $D_{pos}O_{pres}$ in which the position of the objects was disordered and the presentation ordered. Although the estimated and expected probabilities differed in particular in size and not in direction, differences were substantial. One possible explanation for this result is that task $D_{pos}O_{pres}$ differed more than the other tasks from those used by Brainerd and Kingma (1984); these authors used disordered presentation but not disordered position. Therefore, we could not rely on earlier findings when formulating our expectations for this task.

Analysis section 3: relationship of age with verbatim and gist abilities

Based on Brainerd and Kingma (1984, 1985), Reyna and Brainerd (1990), and Reyna (1992) we formulated expectations about the development of the verbatim and

gist abilities. It was expected that the development of verbatim ability is fast and reaches completion at approximately five years of age. Gist ability develops at a slower pace and is not expected to reach full development during childhood (Reyna & Brainerd, 1990; see also Dayton, 1998; Liben & Posnansky, 1977; Marx, 1985a, 1985b; Perner & Mansbridge, 1983; Reyna, 1996; Stevenson, 1972). Therefore, we predicted a positive non-linear relationship between age and both verbatim ability and gist ability. Linear, quadratic and cubic regression curves were fitted to the data to explain the relationships between age (independent variable) and verbatim and gist ability level (dependent variables).

Results

Fig. 9 displays the scatterplots of age and verbatim ability, and age and gist ability. The fit of the linear, quadratic and cubic regression curves did not differ significantly; thus, the curvature of the hypothesized developmental relationships was not supported by the data. The linear model for verbatim ability explained 8% of the variance and that for the gist ability explained 20% .

Discussion

Brainerd and Kingma (1984) investigated the differences in performance between various fixed-age groups, thus ignoring individual differences within age groups. Alternatively, we used regression models to determine the influence of age on ability, and found a low linear correlation between age and verbatim ability and a moderate linear correlation between age and gist ability. Fuzzy trace theory assumes that for transitive reasoning tasks verbatim ability does not improve much after age five; thus, the low correlation between verbatim ability and age might be due to the restricted age range. For preschoolers, a stronger relation might be expected.

Wohlwill (1973, pp. 26–28) and Kessen (1960) argued that chronological age is not a useful variable in statements of functional relationships with behavior because age ignores considerable individual differences in rates of developmental change. Thus, it may be more appropriate to study development by distinguishing groups on the basis of their probability to use verbatim and gist traces instead of fixed-age groups.

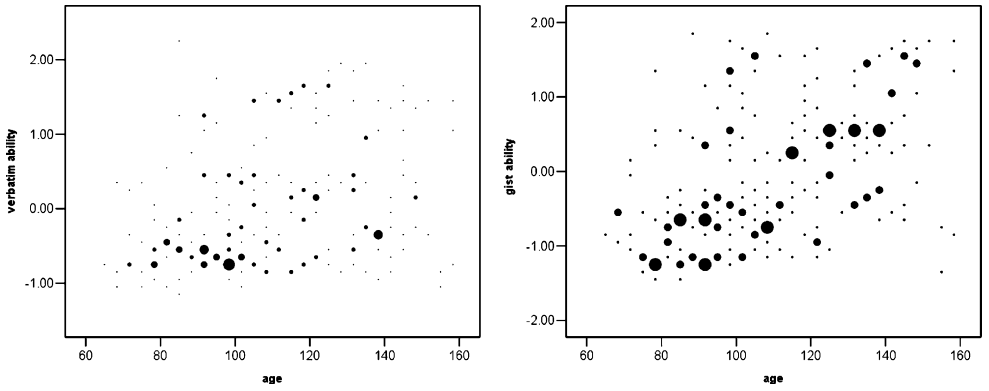


Fig. 9. Scatterplots of verbatim and gist ability scores and age in months (the larger the bullets, the more data points are on the same position).

General discussion

Main findings

Fuzzy trace theory was used to explain individual differences in transitive reasoning. A model was set up in which verbatim and gist-ability levels governed the formation of verbatim and gist traces, and these traces governed performance on memory and transitivity test-pairs. Age was hypothesized to be related to both abilities. A multilevel latent class model was used to handle the dependencies between ability level and trace retrieval, and between trace retrieval and performance on the test pairs. Fitting the model to the data led to two conclusions. First, two abilities had to be distinguished and, second, on the basis of estimated probability structures these abilities could be interpreted as verbatim and gist abilities.

Comparison with earlier findings

Brainerd and Kingma (1984) assumed that both memory and transitivity test-pairs are solved by means of gist traces and were able to show that the unitary-trace model could well explain performance on memory and transitivity test-pairs. Our results agreed with this finding: We also found that children having high gist ability indeed were able to solve both memory and transitivity test-pairs using the appropriate pattern information. For high gist-ability levels and intermediate and high verbatim-ability levels, the unitary-trace model explains both performance on memory and transitivity test-pairs. However, for low or intermediate gist-ability levels the verbatim trace has a strong influence on performance on memory test-pairs, indicating that there is a changing orientation from the use of verbatim traces to both kinds of traces and, finally, to gist traces. For tasks in which the position of the objects is not ordered, as in $D_{\text{pos}}O_{\text{pres}}$ tasks, both high verbatim and gist-trace levels were required to infer the transitive relationship.

Usefulness of the latent class model

The latent class model had several advantages. First, it enabled us to test detailed hypotheses which made it possible to determine which aspects of fuzzy trace theory agreed with observed data and which aspects disagreed. We found that the estimated probabilities for the test-pairs of $O_{\text{pos}}O_{\text{pres}}$ tasks and $O_{\text{pos}}D_{\text{pres}}$ tasks in general agreed with the hypothesized probabilities, but that for $D_{\text{pos}}O_{\text{pres}}$ tasks discrepancies were larger. Indeed, inequality relationships between probabilities were in the hypothesized direction but differences were often greater than expected.

Second, the latent class model enabled us to analyze response patterns on task scores and to predict response processes. Thus, individual differences could be taken into account and latent groups in which children have the same probabilities to use verbatim and gist traces could be identified from the data while probabilities between groups are different. These groups provide an alternative for a priori fixed (age) groups. A more reliable description of differences between these latent groups requires data from even larger numbers of subjects and larger numbers of responses per subject than were used here.

Implications for the debate on developmental stages

The results of this study have implications for the discussion on developmental stages. On the one hand even five-year old children may be able to retrieve high-level gist traces and infer the complete ordering of a task. On the other hand, some twelve-year old children may be more likely to retrieve the lowest trace level and thus do not recognize any ordering in the task. Thus, it is not possible to distinguish clear-cut developmental stages in the development of transitive reasoning (see also Bouwmeester & Sijtsma, in press). Because we used a cross-sectional design, no conclusions could be drawn about the transition from one probability distribution of trace levels to another probability distribution. A longitudinal design is needed for this purpose. This would require an extra level in the multi-level structure to model the dependencies within individual children's data over time.

Appendix A

Let test pairs be indexed by $k = 1,\ldots,7$; tasks by $i = 1,\ldots,12$; and children by $j = 1,\ldots,N$. Response variable $Y_{ijk} = 1$ when child j gives a correct response to test pair k in task i, and $Y_{ijk} = 0$ otherwise. The scores of child j on task i are collected in the vector $\mathbf{Y}_{ij}$, and vector $\mathbf{Y}_j$ denotes the scores of child j on all 12 tasks.

The multilevel latent class model we used contains two ordinal latent variables denoted by X_{ij} and Q_{ij} representing the verbatim and gist traces, respectively, for a particular task i. These two mutually independent latent variables are assumed to have discrete realization between 0 and 1, with equal distances between categories: With three classes per dimension, $x = 0.0$, 0.5, or 1.0, and $q = 0.0$, 0.5, or 1.0. This yields a latent class model with multiple latent variables that Magidson and Vermunt (2001) called a latent class factor model. If we assume that the various tasks performed by a child are independent of one another, the relevant latent class factor model for $\mathbf{Y}_{ij}$ is of the form

$$P(\mathbf{Y}_{ij}) = \sum_x \sum_q P(X_{ij} = x)P(Q_{ij} = q) \prod_{k=1}^{7} P(Y_{ijk}|X_{ij} = x, Q_{ij} = q). \tag{1}$$

This equation reveals the basic assumption of a latent clas model: The scores on the seven test pairs are mutually independent given the latent verbatim and gist-trace levels of child j at task i.

Because of the nesting of tasks within children, the standard assumption of independent observations is not correct for our data. The multiple tasks performed by a child can, however, be assumed to be mutually independent given the child's latent verbatim and gist abilities. These two continuous latent variables, which are denoted by W_j and V_j, respectively, with realization w and v, have the role of random effects in the models for X_{ij} and Q_{ij} (Vermunt, 2003). The abilities or random effects W_j and V_j modify the model for $\mathbf{Y}_{ij}$ described in Eq. (1) as follows:

$$P(\mathbf{Y}_{ij}|W_j = w, V_j = v) = \sum_x \sum_q P(X_{ij} = x|W_j = w)P(Q_{ij} = q|V_j = v) \times \prod_{k=1}^{7} P(Y_{ijk}|X_{ij} = x, Q_{ij} = q). \tag{2}$$

As can be seen, X_{ij} is assumed to depend on W_j, and Q_{ij} on V_j. Moreover, the effects of the continuous latent abilities on the responses are assumed to be fully mediated by the discrete latent trace levels.

The probability associated with all responses of an individual, denoted by $P(\mathbf{Y}_j)$, is obtained by taking the product of $P(\mathbf{Y}_{ij} | W_j = w, V_j = v)$ over the 12 tasks and integrating the two latent ability variables out of the equation. This yields:

$$P(\mathbf{Y}_j) = \int_w \int_v f(W_j = w) f(V_j = v) \left[\prod_{i=1}^{12} P(\mathbf{Y}_{ij} | W_j = w, V_j = v) \right] \mathrm{d}w \, \mathrm{d}v. \tag{3}$$

Note that $P(\mathbf{Y}_{ij} | W_j = w, V_j = v)$ has the form described in Eq. (2), and $f(W_j = w)$ and $f(V_j = v)$ are standard normal univariate distributions.

The three types of model probabilities appearing in Eq. (2) – $P(X_{ij} = x | W_j = w)$, $P(Q_{ij} = q | V_j = v)$, and $P(Y_{ijk} | X_{ij} = x, Q_{ij} = q)$ – are parameterized as logit models. The probability of a correct response of child j on test pair k of task i is restricted by a standard binary logit model of the form

$$P(Y_{ijk} = 1 | X_{ij} = x, Q_{ij} = q) = \frac{\exp(\beta_{0ki} + \beta_{1ki} \cdot x + \beta_{2ki} \cdot q + \beta_{3ki} \cdot x \cdot q)}{1 + \exp(\beta_{0ki} + \beta_{1ki} \cdot x + \beta_{2ki} \cdot q + \beta_{3ki} \cdot x \cdot q)}, \tag{4}$$

where β_{0ki} is an intercept, β_{1ki} and β_{2ki} are the main effects of verbatim-trace level and gist-trace level, respectively, and β_{3ki} is the interaction effect of verbatim and gist-trace levels. The indices k and i indicate that these parameters differ across test pairs and tasks. This is, however, not fully correct since the parameters were restricted to be equal for all four replications of the same task-type (e.g., $\beta_{0k,i+3} = \beta_{0k,i}$). This implies that we have to estimate only three sets of free β parameters.

The other two parts of the model, capturing the relative sizes of the verbatim and gist-trace levels given the verbatim and gist-ability levels, are modeled as

$$P(X_{ij} = x | W_j = w) = \frac{\exp(\gamma_{0x} + \gamma_1 \cdot x \cdot w)}{\sum_x \exp(\gamma_{0x} + \gamma_1 \cdot x \cdot w)},$$

and

$$P(Q_{ij} = q | V_j = v) = \frac{\exp(\gamma_{2q} + \gamma_3 \cdot q \cdot v)}{\sum_q \exp(\gamma_{2q} + \gamma_3 \cdot q \cdot v)}.$$

These are adjacent-category ordinal logit models similar to the ones used in partial-credit models, which are item response models for ordinal items. The γ parameters are assumed to be equal across the 12 tasks.

The multilevel latent class models were estimated by means of maximum likelihood using an adapted version of the EM algorithm (Vermunt, 2003, 2004). This procedure is implemented in version 4.0 of Latent Gold (Vermunt & Magidson, 2005), a Windows-based program for latent class analysis, that is available at www.statisticalinnovations.com.

References

Adams, M. J. (1978). Logical competence and transitive inference in young children. *Journal of Experimental Child Psychology, 25*, 477–489.

Berch, D. B. (1979). Coding of spatial and temporal information in episodic memory. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior*. New York: Academic Press.

Bouwmeester, S., & Aalbers, T. (2004). TRANRED 2. Tilburg: Tilburg University.

Bouwmeester, S., & Sijtsma, K. Latent class modeling of phases in the development of transitive reasoning. *Multivariate Behavioral Research*, in press.

Bower, G. H. (1971). Adaption-level coding of stimuli and serial position effects. In M. H. Appley (Ed.), *Adaption-level theory*. New York: Academic Press.

Braine, M. D. S. (1959). The onthogeny of certain logical operations: Piaget's formulation examined by nonverbal methods. *Monographs for the Society for Research in Child Development, 27*, 41–63.

Brainerd, C. J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy-trace theory of transitivity development. *Developmental Review, 4*, 311–377.

Brainerd, C. J., & Kingma, J. (1985). On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology, 17*, 210–247.

Brainerd, C. J., & Reyna, V. F. (1990a). Gist is the grist: fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*, 3–47.

Brainerd, C. J., & Reyna, V. F. (1990b). Inclusion illusions: fuzzy-trace theory and perceptual salience effects in cognitive development. *Developmental Review, 10*, 365–403.

Brainerd, C. J., & Reyna, V. F. (1992). The memory independence effect: what do the data show? what do the theories claim? *Developmental Review, 12*, 164–186.

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100*, 42–67.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review, 24*, 396–439.

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature, 232*, 456–458.

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage in the development of transitivity. *Developmental Psychology, 24*, 542–551.

Chapman, M., & Lindenberger, U. (1992). Transitivity judgments, memory for premises, and models of children's reasoning. *Developmental Review, 12*, 124–163.

Clogg, C. C. (1988). Latent class models for measuring. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 137–205). New York: Plenum Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cooney, J. B. (1995). Fuzzy-trace theory: a proposed test of individual differences. *Learning and Individual Differences, 7*, 139–144.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Flavell, J. H. (1970). Stage-related properties of cognitive development. *Cognitive Psychology, 2*, 421–453.

Grieve, R., & Nesdale, A. R. (1979). Observations on a test of transitive inference in children. *Australian Journal of Psychology, 31*, 43–48.

Halford, G. S. (1979). Measurement and memory in transitivity: a reply to Grieve and Nesdale. *Australian Journal of Psychology, 31*, 49–56.

Halford, G. S., & Galloway, W. (1977). Children who fail to make transitive inferences can remember comparisons. *Australian Journal of Psychology, 29*, 1–5.

Heinen, T. (1996). *Latent class and discrete latent trait models*. Thousand Oaks, CA: Sage.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Kessen, W. (1960). Research design in the study of developmental problems. In P. H. Mussen (Ed.), *Handbook of research methods in child development* (pp. 257–262). New York: Wiley.

Kingma, J. (1981). De ontwikkeling van quantitatieve en relationele begrippen bij kinderen van 4 tot 12 jaar (The development of quantitative and relational concepts in 4- to 12-year-old children). Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Liben, E. H., & Posnansky, C. J. (1977). Inference on inference. The effects of age, transitive ability, memory load, and lexical factors. *Child Development, 48*, 490–497.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology, 31*, 223–264.

Marx, M. H. (1985a). Retrospectives reports on frequency judgments. *Bulletin of the Psychonomic Society, 23*, 309–310.

Marx, M. H. (1985b). More retrospective reports on event-frequency judgments: shift from multiple traces to strength factor with age. *Bulletin of the Psychonomic Society, 24*, 183–185.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks, CA: Sage.
Perner, J., & Mansbridge, D. G. (1983). Developmental differences in encoding. *Child Development, 54*, 710–719.
Perner, J., Steiner, G., & Staehelin, C. (1981). Mental representation of length and weight series and transitive inferences in young children. *Journal of Experimental Child Psychology, 31*, 177–192.
Piaget, J. (1942). *Classes, relations et nombres: essai sur les groupement logistique et sur la réversibilité de la pensée*. Paris: Collin.
Piaget, J. (1947). *La psychologie de l'intelligence*. Paris: Collin.
Piaget, J., Inhelder, B., & Szeminska, A. (1948). *La géométric spontanée de l'enfant*. Paris: Presses Universitaires de France.
Potts, G. R. (1972). Information processing strategies used in the encoding of linear orderings. *Journal of Verbal Learning and Verbal Behavior, 11*, 727–740.
Raijmakers, M. E. J., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review, 24*, 289–321.
Reyna, V. F. (1992). Reasoning, remembering, and their relationship: social, cognitive, and developmental issues. In M. L. Howe, C. J. Brainerd, & V. F. Reyna (Eds.), *Development of long-term retention* (pp. 103–132). New York: Springer-Verlag.
Reyna, V. F. (1996). Conceptions of memory development with implications for reasoning and decision making. *Annals of Child Development, 12*, 87–118.
Reyna, V. F., & Brainerd, C. J. (1990). Fuzzy processing in transitivity development. *Annals of Operations Research, 23*, 37–63.
Reyna, V. F., & Brainerd, C. J. (1992). A fuzzy-trace theory of reasoning and remembering: paradoxes, patterns, and parallelism. In A. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes vol. 2* (pp. 235–259). Hillsdale, NJ: Erlbaum.
Reyna, V. F., & Brainerd, C. J. (1995a). Fuzzy-trace theory: an interim synthesis. *Learning and Individual Differences, 7*, 1–75.
Reyna, V. F., & Brainerd, C. J. (1995b). Fuzzy-trace theory: some foundational issues. *Learning and Individual Differences, 7*, 145–162.
Riley, C. A. (1976). The representation of comparative relations and the transitive inference task. *Journal of Experimental Child Psychology, 22*, 1–22.
Riley, C. A., & Trabasso, T. (1974). Comparatives, logical structures, and encoding in a transitive inference task. *Journal of Experimental Child Psychology, 17*, 187–203.
Rindskopf, D. (1987). Using latent class analysis to test developmental models. *Developmental Review, 7*, 66–85.
Russell, J. (1981). Children's memory for premises in a transitive measurement task assessed by elicited and spontaneous justifications. *Journal of Experimental Child Psychology, 31*, 300–309.
Smedslund, J. (1963). Development of concrete transitivity of length in children. *Child Development, 34*, 389–405.
Smedslund, J. (1965). The development of transitivity of length: a comment on Braine's reply. *Child Development, 36*, 577–580.
Smedslund, J. (1969). Psychological diagnostics. *Psychological Bulletin, 71*, 237–248.
Sternberg, R. J. (1980). The development of linear syllogistic reasoning. *Journal of Experimental Child Psychology, 29*, 340–356.
Sternberg, R. J., & Weil, E. M. (1980). An aptitude × strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72*, 226–239.
Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.
Stevenson, H. W. (1972). *Children's learning*. New York: Appleton-Century-Crofts.
Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R. V. Kail, J. W. Hagen, & J. M. Belmont (Eds.), *Perspectives on the development of memory and cognition* (pp. 333–366). Hillsdale, NJ: Erlbaum.
Trabasso, T., & Riley, C. A. (1975). The construction and use of representations involving linear order. In R. L. Soslo (Ed.), *Information processing and cognition: The Loyola symposium*. Hillsdale, NJ: Erlbaum.
Trabasso, T., Riley, C. A., & Wilson, E. G. (1975). The representation of linear order and spatial strategies in reasoning: a developmental study. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 201–229). Hillsdale, NJ: Erlbaum.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519–538.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213–239.

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric hierarchical nonlinear models. *Statistica Neerlandica, 58*, 220–233.

Vermunt, J. K., & Magidson, J. (2005). *Latent Gold 4.0*. Belmont, MA: Statistical Innovations Inc..

Verweij, A.C. (1994). *Scaling transitive inference in 7–12 year old children*. Unpublished doctoral dissertation, Vrije Universiteit Amsterdam, The Netherlands.

Wohlwill, J. F. (1973). *The study of behavioral development*. New York: Academic Press.